Using Belief Networks to Enhance Sharing of Medical Knowledge Between Sites With Variations in Data Accuracy

William R. Hogan, Department of Medicine
Michael M. Wagner, Section of Medical Informatics
University of Pittsburgh

Differences in data definition between sites are a known obstacle to sharing of reminder-system rule sets. We identify another data characteristic—data accuracy—with implications for sharing. We reviewed the literature on data accuracy and found reports of high error rates for many data classes used by reminder systems (e.g., problem lists). The accuracy of other, equally important, data classes had not been characterized. Wide variations in accuracy between sites has been observed, suggesting that such differences may pose a previously unrecognized barrier to sharing of reminder rules. We propose a belief-network model for encoding reminder rules that explicitly models site-specific data accuracy and we discuss how encoding knowledge in this format may lower the cost and effort required to share reminder rules between sites.

INTRODUCTION

Knowledge bases (KBs) for reminder systems are costly to develop and maintain, and yet cover only a limited fraction of the domain of medicine. For these reasons, sharing of KBs is a significant research area.

Researchers have identified local differences in *data definition* as an obstacle to sharing KBs among sites (1, 2). As we shall discuss, differences in *data accuracy* may pose an additional obstacle to sharing. In this paper, we review the literature on data accuracy with particular emphasis on those aspects relevant to sharing, and we propose a belief-network model that may reduce the cost of adapting rules developed at one site for use in another site.

DATA ACCURACY

Medical data are defined and collected with a marked degree of variability and inaccuracy. The taking of a medical history, the performance of a physical examination, the interpretation of laboratory tests, even the definition of diseases, are surprisingly inexact. We consider the implications of this reality for computerized medical information systems, quantitative techniques for medical diagnosis, and the evaluation of bioengineering technology. Komaroff (3).

Because we are primarily interested in reminder systems, which draw inferences about patients from data in electronic medical records (EMRs), we define *accurate patient data* to be data that represents the true state of the patient. There are at least three ways that data in an EMR may become inaccurate under our definition. First, EMRs receive observations and data about patients from diverse sources (Fig. 1). Data may be entered directly by patients, via laboratory systems, or via pathways emanating from physician observations, including transcription, data-entry personnel, and direct clinician entry. Inaccuracies of patient representation may be introduced at each point in this flow of information from patient to record. Second, the true state of a patient changes with time due to the effects of disease or treatments; data error may accumulate from a lack of recent observations. For instance, a medication (e.g. warfarin) may be stopped between visits. If this change were not captured and a reminder system were to evaluate its rule set after this change (e.g. triggered by a scheduled visit or the passage of time), it would do so on the basis of inaccurate data. Third, due to representational limitations of the EMR, it may not be possible to represent some patient characteristics.

Studies that have examined *overall* data accuracy (in research and epidemiologic databases as well as clinical information systems), have found data accuracy rates that range from 57 to 96% (Table 2). Data accuracy for *diagnoses* ranged from 33.7% for a diagnosis of smoking to 98.9% for anemia (another author reported 54.1% for anemia). Data accuracy for procedures ranged from 30% of events captured and accurately recorded (suturing of lacerations during delivery) to 97% (Caesarean-section). Accuracy for surgical complications were 50% to 95.7%. In one study, descriptive modifiers from dictated endoscopy and ultrasonography reports had accuracy rates of 86.4% to 95.7%. For several classes of data commonly referred to in reminder rules—laboratory test results, allergies, diagnoses, and, until recently, medications—there has been no characterization of data accuracy.

Of particular importance to sharing is one study that compared accuracy of the same data classes at similar district general hospitals in southern England

0195-4210/95/$5.00

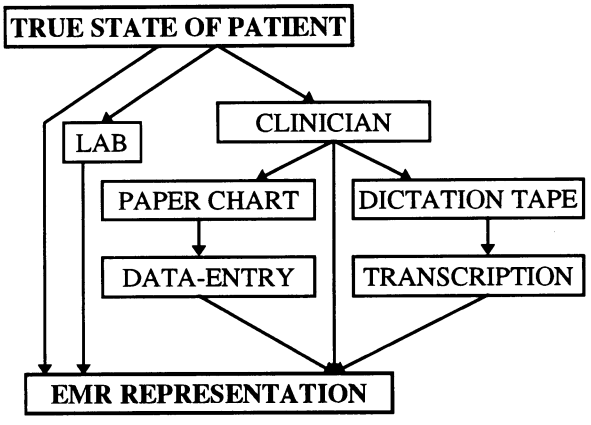

Fig. 1 Flow of information from the patient to EMR.

(4). These two hospitals were located in the *same* part of the country, had *similar* staff, used their systems for the *same* period of time, and used the *same software package* (the Manchester Orthopaedic Database). The study examined the accuracy of dictated *keywords* in systems used by the Department of Orthopedics at each hospital (departments A and B). Keywords described the clinical details of patient care and were of three types: diagnosis, procedure, and complication. The authors measured the rates of both inaccurate and missing keywords (keywords that should have been used) for 100 sequential inpatient admissions at each department. The keywords in the computer system were compared to an 'ideal' set of keywords generated by manual review of medical records. The overall accuracy at department B was 88%, with a 25% keyword omission rate. The overall accuracy at department A was 57% with a 55% omission rate. For diagnoses, the accuracy was 67% for A (47% omission) and 91% for B (26% omission). For procedures, accuracy was 44% for A (47% omission) and 86% for B (20% omission). For complications, accuracy was 50% for A (83% omission) and 77% for B (44% omission). The differences achieved statistical significance for overall rate, and rates for diagnosis and procedure keywords. Thus, despite their numerous similarities (setting, time of use, software, staffing), these two systems had significantly different rates of accuracy.

From our review, we conclude that error rates are high, the accuracy of many data classes typically used by reminder systems has not been characterized (although some of these data classes, such as laboratory tests, are collected automatically and thus are likely to be accurate), and accuracy varies widely among sites that potentially may want to share KBs.

IMPLICATIONS FOR SHARING

Reminder systems are developed and function in the context of inaccurate EMR data. They sometimes send false alarms triggered by inaccurate data, but the developers and recipients of such reminders understand that the systems have this property. Problems may arise, however if we move rules developed at one location to another with *higher* levels of data error. (There are many other differences between sites that can cause problems in portability. In this paper, we are thinking of two sites, as in the English study, that differ mainly in data accuracy).

If data inaccuracy could be eliminated, it would not be a barrier to sharing. However, this is not likely to happen soon. People will incorrectly record observations, social forces may introduce distortion (e.g., diminishing record confidentiality may cause physicians to withhold or falsify information in the medical record (5)), and the dynamic nature of patient health-states and the intermittent capture of medical data by EMR systems, make a perfect representation of a patient virtually unattainable.

RULE PORTABILITY

The objective of sharing is to reduce the cost and effort required to implement and maintain knowledge-based systems.

The idea of sharing can mean something as simple as communicating which practice guidelines (PGs) have been successfully implemented to transferring the implemented PGs in a syntax that the receiving system could interpret. It is not clear from research to date that transferring implemented rules or PGs reduces the cost or effort for the receiving site more than communicating the topics of reminding.

To appreciate the role that data accuracy plays in sharing (and to understand why rules are not portable, in the general case), consider the meaning of a rule at the receiving site. At the receiving site, the meaning of the elements in the rule is ambiguous. The receiving site must choose between interpreting the elements as references to patient characteristics, or as references to database variables. Furthermore, there is no characterization of the *accuracy* of the rule itself, which we define as the precision with which it identifies patients for whom the reminder is appropriate. Consider, for example, a hypothetical PG that states that we should obtain a screening cholesterol for 40 year-old male smokers. This PG is an imperfect representation of how an expert physician would practice (a cardiologist would not check a cholesterol if such a patient were moribund). If we implement the PG in a reminder system,

Table 2. Reported data accuracy rates from the literature

Data Class	References	Accuracy Rates
All (overall accuracy)	Aas, 1988(6); Barrie, 1992 (7); Basden, 1979 (8); Jelovsek, 1978 (9); Lloyd, 1985 (10); Payne, 1991 (11); Rao, 1986 (12); Ricketts, 1993 (4); Teperi, 1993 (13); Wilton, 1993 (14)	65.5%[a], 96%[b], 93%[c], 94%[d], 78%[e], 96%[f], 95.8%[g], 57%-88%[h], 95%, 94.1%
Diagnosis/problem list	Aas, 1988 (6)	88.5%[a]
	Barrie, 1992 (7)	96.5%[b]
	Block, 1989 (15)	33.7- 92.3%[i]
	Rao, 1986 (12)	98.3- 98.9%[j,g]
	Roos, 1993 (16)	95%
	Teppo, 1994 (17)	66%-90%[k, q]
	McGonigal, 1992 (18)	41.3%-93%[l]
Procedures/Operations	Aas, 1988 (6); Barrie, 1992 (7); Roos, 1993 (16); Skinner, 1988 (19); Teperi, 1993 (13)	89.2%[a], 82%[b], 90%[m], 70-95%[n, q], 30-97%[o, q]
Surgical Complications	Aas, 1988 (6); Barrie, 1992 (7); Ricketts, 1993 (4)	95.7%[a], 54.1%[b], 50%-77%[h]
Modifiers of pathology findings	Kuhn, 1991 (20)	86.4%-95.7%[p, q]
Medications	Block, 1995(21); Wagner, 1995 (22)	70.2%[r], 77.6%[r]

[a] refers to % of records with no inaccuracies in the data used to determine Diagnosis Related Groups
[b] also reports missing data, with 38% of 'keywords' for diagnoses, procedures, and complications missing
[c] refers to % of records without a "major error"
[d] median rate over 98 variables (range 16-100%), includes only errors of omission
[e] out of 20,832 data items that were included in automated discharge abstracts for 1829 records
[f] this study reports that 10% of patients were omitted from an immunization database
[g] accuracy rate not specifically calculated or stated by author
[h] rates for two hospitals using same system, see text for discussion of this study
[i] smoking–33.7%, pediatric anemia–35%, adult anemia–54.1%, urinary tract infection–54.8%, and pregnancy–92.3%
[j] refer to anemia—98.9%, coma—98.6%, and 'pyrexia under observation'—98.3%
[k] varying rates dependent on type of cancer
[l] 41.3% for presenile dementia (17.3% omission), 93% for presenile Alzheimer's disease (6.6% omission)
[m] for 8 of 11 procedures (overall rate for all 11 not given)
[n] major operations-70%, all orthopedic operations-95%, and hip replacements-84%
[o] best was C-section-97%, worst was suturing of lacerations-30%
[p] refers to omissions of descriptive modifiers of gut pathology found on endoscopy and ultrasonography
[q] includes only errors of omission
[r] data on medication accuracy is currently unpublished

additional imprecision is introduced by mismatches between the PG and the variables in the EMR, due to differences in both data definition and data accuracy.

At the original site, knowledge engineers have come to terms with the imprecision of their rules. Based on data about (or subjective estimates of) the rates of false and true alarms and their utilities, they may have made a series of decisions about the form of the rule, and whether to implement it at all. Currently, there is no way—other than verbal communication—for them to convey this information to a receiving site. The receiving site, therefore, must discover for itself if adding the rule to its system is of net positive benefit. The fact that logical languages cannot represent rates of false alarms (or their utilities) is a fundamental obstacle to sharing. For this reason, we propose the following probabilistic formulation of reminder systems and discuss how it may reduce the work involved in sharing reminder rules.

BELIEF-NETWORK REMINDER RULES

The decision whether to incorporate a shared reminder rule into a reminder system should depend on the expected utility of the rule at the new site. The expected utility of a rule is determined by the expected rate of false and true positives at the receiving site, and the utility of those events. Fig. 2 is a belief-network representation of the relationship between data accuracy and false and true alarms. (We do not discuss the utility of these alarms in this paper. Conveying information that allows the receiving site to estimate the rate of true and false alarms that they will experience, however, may reduce the cost of sharing, as we shall discuss.)

The box in Fig. 2 identifies a subgraph of the belief network that represents the PG *if a patient has peptic ulcer disease and she is taking a non-steroidal anti-inflammatory drug, then we should stop the non-steroidal drug*. We refer to the nodes in the belief network that represent PG preconditions (i.e., *peptic ulcer disease* and *non-steroidal drug use)* as **patient-state** nodes because a rule author usually means the true state of the patient when she formulates a rule. We use capital letters for patient-state variables. The **PG preconditions satisfied** node (labeled *Pg sat*) represents whether the PG preconditions are satisfied, identical to whether the logic of the rule is satisfied.

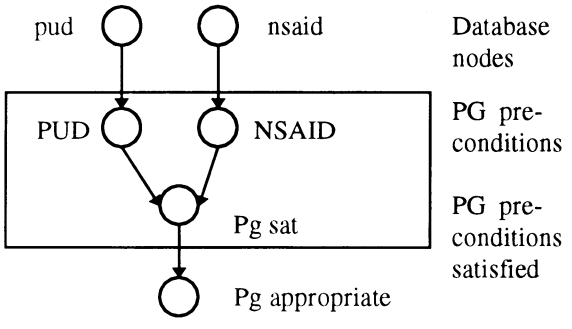

Fig. 2. Belief-network representation of the reminding rule *if a patient is taking a non-steroidal anti-inflammatory drug (NSAID) and the patient has peptic ulcer disease (PUD), then stop the NSAID.*

By specifying the following conditional probability distribution for the belief network, the boxed section of the belief network will mimic the PG.

P(Pg sat = true | NSAID = true, PUD = true) = 1.0
P(Pg sat = true | NSAID = true, PUD = false) = 0.0
P(Pg sat = true | NSAID = false, PUD = true) = 0.0
P(Pg sat = true | NSAID = false, PUD = false) = 0.0
P(Pg sat = false | NSAID = true, PUD = true) = 0.0
P(Pg sat = false | NSAID = true, PUD = false) = 1.0
P(Pg sat = false | NSAID = false, PUD = true) = 1.0
P(Pg sat = false | NSAID = false, PUD = false) = 1.0

We assume that PG preconditions refer to patient states, not database representations, therefore we distinguish between database variables and patient-state variables. We use lower-case names to refer to database variables. We represent the differences between database variables and the patient-state variables using conditional probability distributions, for example, *P(PUD | pud)* and *P(NSAID | nsaid)*. These distributions describe the data accuracy of the database variables.

We also represent explicitly the difference between a patient who matches the logic of a PG, and a patient who expert physicians would agree matches the intent of the PG (i.e., good medical practice) by adding a variable to represent whether the PG is truly appropriate. The conditional probability distribution *P(Pg appropriate | Pg sat)* characterizes the fidelity of the PG to ideal medical practice.

In this model, the conditional probability distributions that model the logic of the rules would not change from site to site, but the other distributions could change markedly. For example, the probability distributions for *P(NSAID | nsaid)* and *P(Pg appropriate | Pg sat)* may differ from site to site due to differences in data accuracy, and differences in context, respectively.

To see how our approach may improve the efficiency with which rule sets are imported into a new setting, consider that, with existing rules, the receiving site receives only what is contained within the box in Fig. 2 (possibly accompanied by a history of the rule, and references as in the Arden specification). In our new model, the original site would send the complete model, including the probability distributions that define its data accuracy, e.g., *P(NSAID | nsaid),* and any data about *P(Pg appropriate | Pg sat).* The receiving site could, using its own accuracy data, use the belief network to compute *P(Pg sat | nsaid, pud)* which provides a rough estimate of the true and false alarm rates that it could expect. It could compare this value against the level at which the sending site was operating (we are assuming that the sending site has similar utilities for true and false alarms, and that they are sending a rule found to be acceptable in their setting). If the expected true alarm rate for the receiving site were higher, and the false alarm rate lower, the receiving site would have some measure of confidence that the rule would work in its setting. If the pattern were reversed, it would have to either consider whether to implement the rule at all, attempt to improve data accuracy, try to modify the rule to minimize false alarms, or empirically test the rule itself.

If the receiving site were also given the sending site's *P(Pg appropriate | Pg sat)* and believed this to be similar to its own, it could estimate *P(Pg appropriate | nsaid, pud),* forming an even more accurate estimate of the number of false and true alarms it would likely experience. It could use this information to prioritize its efforts, if for example, it were importing many rules, or had limited resources.

DISCUSSION

Sharing is a potential means to reduce the cost of medical KB development. There are many obstacles to sharing, however, including site differences in *data definition* and *data accuracy*.

Our critical examination of prior studies on data accuracy revealed that data accuracy varies

considerably from site to site, error rates are often high, and the accuracy of some data classes typically used by reminder systems has not been characterized.

Rules transferred between sites with variations in data accuracy can be expected to produce unpredictable rates of false and true alarms at the new site. In some cases these differences will be sufficient to change the net expected benefit of implementing the rule from positive to negative. Knowledge base developers currently have few tools to diagnose such problems before they occur, hence they must evaluate the merit of every imported rule.

We proposed a general belief-network representation of reminder rules that distinguishes between a patient's true state and the database representation of that state. It also distinguishes between the probability that a PG's preconditions are met, and the probability that the PG is truly appropriate. If a receiving site has characterized its data accuracy, this model will enable it to estimate the number of false alarms and true alarms to expect when importing reminder rules.

The utility of such alarms might also impact on the decision to import a rule, and this information could also be conveyed to the receiving site. Author *mmw* has discussed how *utility* of reminders might be measured and used to provide a rational basis for setting the optimal rates of true and false alarms (23).

We have demonstrated how to represent explicitly data error in a reminder system using a probabilistic formalism that conditions the true state of a patient on the data contained in an EMR. There are two additional advantages of our belief-network reminder system. First, it is tolerant of PG parameters that may be missing from an EMR. If, in our example, the PG states that we should not send a reminder if a patient is on cytotec for prophylaxis against NSAID-related PUD, but whether a patient is on cytotec or not is not represented at the receiving site (perhaps it is not in the formulary), the belief network can still be evaluated with the patient-state node representing cytotec in an indeterminate state. Second, this model can represent differences in data definition. If the patient-state nodes of the PG are defined using a common medical vocabulary, the receiving site could represent the semantic difference between its database variables and the PG concept as a conditional probability.

References

1. Clayton PD, Pryor TA, Wigertz OB, Hripcsak G. Issues and Structures for Sharing Medical Knowledge among Decision-Making Systems: The 1989 Arden Homestead Retreat. Proc 13th SCAMC 1989:116-121.
2. Pryor TA, Hripcsak G. Sharing MLM's: An Experiment between Columbia-Presbyterian and LDS Hospital. Proc 17th SCAMC 1993:399-403.
3. Komaroff AL. The Variability and Inaccuracy of Medical Data. Proc of the IEEE 1979;67(9):1196-1207.
4. Ricketts D, Patterson M, Newey M, Hitchin D, Fowler S. Markers of data quality in computer audit: the Manchester Orthopaedic Database. Ann R Coll Surg Engl 1993;75(6):393-396.
5. Burnum JF. The Misinformation Era: The Fall of the Medical Record. Ann Int Med 1989;110(6):482-484.
6. Aas IHM. Quality of Hospital Data and DRGs. Scand J Soc Med 1988;16(4):223-226.
7. Barrie JL, Marsh DR. Quality of data in the Manchester orthopaedic database. BMJ 1992;304(6820):159-162.
8. Basden A, Clark EM. Errors in a computerized medical record system. Med Inform 1979;4(4):203-208.
9. Jelovsek F, Hammond W. Formal Error Rate in a Computerized Obstetric Medical Record. Meth Inform Med 1978;17(3):151-157.
10. Lloyd SS, Rissing JP. Physician and Coding Errors in Patient Records. JAMA 1985;254(10):1330-1336.
11. Payne T, Kanvik S, Seward R, et al. Development of an Immunization Tracking System in a Large Health Maintenance Organization. Proc 15th SCAMC 1991:131-135.
12. Rao NG, Prasad BE, Guha SK, Reddy PG. A user-oriented validation method for clinical data. Med Inform 1986;11(4):317-328.
13. Teperi J. Multi method approach to the assessment of data quality in the Finnish Medical Birth Registry. J Epidemiol Community Health 1993;47(3):242-247.
14. Wilton R, Pennisi AJ. Evaluating the Accuracy of Transcribed Clinical Data. Proc 17th SCAMC 1993:279-283.
15. Block B, Brennan JA. Reliability of Morbidity Data in a Computerized Medical Record System. Proc 8th Ann AAMSI Cong 1989:21-30.
16. Roos LL, Mustard CA, Nicol JP, et al. Registries and Administrative Data: Organization and Accuracy. Med Care 1993;31(3):201-212.
17. Teppo L, Pukkala E, Lehtonen M. Data Quality and Quality Control of a Population-Based Cancer Registry. Acta Oncol 1994;33(4):365-369.
18. McGonigal G, McQuade C, Thomas B. Accuracy and Completeness of Scottish Mental Hospital In-patient Data. Health Bulletin 1992;50(4):309-314.
19. Skinner PW, Riley D, Thomas EM. Use and abuse of performance indicators. BMJ 1988;297:1256-1259.
20. Kuhn K, Swobodnik W, Johannes RS, et al. The Quality of Gastroenterological Reports Based on Free Text Dictation: An Evaluation in Endoscopy and Ultrasonography. Endoscopy 1991;23(5):262-264.
21. Block B. Unpublished Data, 1995.
22. Wagner MM, Hogan WR. The Accuracy of Outpatient Medication Data in Electronic Medical Records. Report SMI-95-04, U. Pittsburgh, 1995.
23. Wagner MM, Cooper, GF. Decision-Theoretic Information Pretrieval: A Generalization of Reminding. Proc 17th SCAMC 1993:512-516.